D294–D296 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki038

The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis

J. R. Cole[1,*], B. Chai[1], R. J. Farris[1], Q. Wang[1], S. A. Kulam[1], D. M. McGarrell[1], G. M. Garrity[1,2] and J. M. Tiedje[1,2]

[1]Center for Microbial Ecology and [2]Department of Microbiology and Molecular Genetics, Biomedical Physical Sciences, Michigan State University, East Lansing, MI 48824-4320, USA

Received September 14, 2004; Revised and Accepted September 24, 2004

ABSTRACT

The Ribosomal Database Project (RDP-II) provides the research community with aligned and annotated rRNA gene sequences, along with analysis services and a phylogenetically consistent taxonomic framework for these data. Updated monthly, these services are made available through the RDP-II website (http://rdp.cme.msu.edu/). RDP-II release 9.21 (August 2004) contains 101 632 bacterial small subunit rRNA gene sequences in aligned and annotated format. High-throughput tools for initial taxonomic placement, identification of related sequences, probe and primer testing, data navigation and subalignment download are provided. The RDP-II email address for questions or comments is rdpstaff@msu.edu.

DESCRIPTION

Release 9 introduces substantial changes to the Ribosomal Database Project (RDP). These changes are in response to the rapidly increasing number of available ribosomal RNA gene sequences (rRNA sequences) and the trend toward high-throughput rRNA sequencing with the concomitant need for high volume rRNA analysis tools. This paper describes changes since the 2003 description (1). Details about the data and analysis services can be found at the RDP-II website (http://rdp.cme.msu.edu/).

Sequences. The RDP obtains bacterial rRNA sequences from the International Nucleotide Sequence Databases (INSD: GenBank/EMBL/DDBJ) on a monthly basis. These sequences are aligned against a general bacterial rRNA model using a modified version of RNACAD (2), a Stochastic Context Free Grammar (SCFG)-based rRNA aligner that directly incorporates rRNA secondary structure information into its internal model. This aligner is trained on a set of high-quality hand-aligned sequences and incorporates the conserved bacterial secondary structure model of Gutell and co-workers (3). As of release 9.21 (August 2004), the database contained 101 632 total small subunit bacterial rRNA sequences. Of these, 39 772 were near full-length (⩾1200 bases), 54 316 came from uncultured organisms and 4431 were from type strains of validly named bacterial species.

Taxonomy. All Release 9 tools use a new hierarchical framework (RDP Hierarchy) differing significantly from the hierarchy provided with previous RDP releases. The RDP Hierarchy is based on the new phylogenetically consistent higher-order bacterial taxonomy proposed by Garrity *et al*. (4) (http://dx.doi.org/10.1007/bergeysoutline). This hierarchy provides order to the collection. It provides a phylogenetic framework into which to place results of the RDP analysis functions, and it provides an entry point for users looking for sequences from specific groups of organisms. New sequences are placed into the RDP Hierarchy using the RDP Classifier (see below).

Analysis services

The RDP analysis services have been completely revised to support the emerging trend toward high-throughput rRNA sequence analysis in microbial ecology and related disciplines. Three of the tools listed below incorporate the concept of data filters. The user can choose to apply up to three data filters on the view or analysis. By applying the three filters, the user can (i) include only environmental clone or only isolate sequences; (ii) include only sequences ⩾1200 bases in length (near full-length) or only shorter sequences; and (iii) include only sequences from type strains or only non-type strain sequences. The latter filter is of special importance since type strains act as a link between rRNA-based phylogeny and taxonomy. A more detailed description of each analysis service can be found at the RDP website.

*To whom correspondence should be addressed. Tel: +1 517 432 4998; Fax: +1 517 353 8957; Email: rdpstaff@msu.edu

Hierarchy Browser allows rapid navigation through the RDP sequence data. The browser presents views of the RDP sequences placed either in the RDP Hierarchy, or optionally in the NCBI taxonomy hierarchy (5). While navigating, the browser automatically expands an appropriate number of hierarchical levels to fit the display. At any time, the user can select for later download of both individual sequences and those of entire taxa. Data filters can be applied at any time to limit the display to specific data subsets. In addition, the user can quickly search for words or phrases in the sequence definition line. This includes the organism name and strain designation (if available), culture collection identifiers and INSD nucleic acid accession identifiers.

RDP Classifier places sequences into the RDP Hierarchy. Optimized for large query sets, it can be used to give an initial taxonomic placement for a single sequence or hundreds of sequences. The first result page summarizes the assignments on an interactive display similar to that of the Hierarchy Browser. Each node in the hierarchy lists the number of user queries assigned to that taxonomic rank. A confidence estimate is generated for each assignment, and the assignments are displayed only when the estimate is above a user-specified confidence threshold. At any time, the user can switch to a detail view showing the detailed taxonomic assignments and confidence scores for any subset of query sequences. These assignment details can also be downloaded in a file suitable for import into popular spreadsheet programs.

Sequence Match is a complete re-implementation of the original Sequence Match method (1). Sequence Match finds sequences similar to a user's query sequences using a word matching strategy not requiring prior alignment. Sequence Match is more accurate than BLAST (6) at finding closely related rRNA sequences (Table 1). The related sequences returned by Sequence Match serve as a good starting point for more detailed examination of relatedness by classical phylogenetic or other methods. The initial result page presents a *k*-nearest neighbor (*k*-NN) classifier assignment of the query sequences. A query is assigned to the lowest taxonomic rank that includes the *k* highest scoring database sequences. The value of *k*, as well as the three data filters can be changed at will in this view. The user can switch from the summary *k*-NN view to a detailed results view for any query sequence. In this view, the top *k* database matches to the query are displayed in the RDP Hierarchy. In this mode, any subset of the matches can be selected for transfer to the Hierarchy Browser and later download. A third view presents sets of results in a format suitable for download.

Probe Match is a complete re-implementation of our previous Probe Match program (1). It uses a more efficient algorithm that is better suited to the amount of rRNA data available today and in the foreseeable future. The new Probe Match accepts a candidate primer/probe, optionally with ambiguity codons, of up to 64 bases in length. While our previous version searched for hits within a specified number of mismatches (Hamming distance), the new version finds hits with a combination of mismatches and insertion/deletions (edit distance). Since some single insertion/deletion may be no more deleterious than a single mismatch, this new capability offers a significant improvement in the detection of potential cross-hybridization. In our previous implementation, the high percentage of partial sequences in the database limited the program's utility; it was difficult to determine if database entries failed to match simply because the sequence was incomplete in the target region. In this new version, the users can restrict analysis to database entries containing sequence data for the candidate probe target region of the rRNA molecule. (However, the search is not limited to this region of the molecule.) Similar to the other new programs, the results are displayed in an interactive version of the RDP Hierarchy. Each taxonomic rank lists the total number of sequences searched and the number matching within a user-specified edit distance. This maximum edit distance can be changed on the fly. For any hierarchy node, users can switch to a detail view listing the matching sequences. A third format is suitable for download and import into spreadsheet or other programs.

Table 1. rRNA search performance

Program[a]	Percentage of 16S rRNA queries[b] returning the most similar sequence[c] among the highest scoring *N* results		
	N = 1	*N* = 10	*N* = 20
Sequence Match	65	92	95
BLAST	39	53	55

[a]For both programs, the dataset consisted of 37 456 near full-length (≥1200 base) rRNA sequences from the RDP release 9.20 alignment database.
[b]Query sequences (1000) were selected at random from the dataset.
[c]The most similar sequence to each query was determined by exhaustive pairwise similarity comparison of each query against the dataset. In cases of a tie in pairwise similarity, we required only one of the ties to be returned by the program.

RDP-II ACCESS AND CONTACT

The RDP-II data and analysis services can be found at http://rdp.cme.msu.edu/. The RDP's mission includes user support. The address for email support is rdpstaff@msu.edu. Telephone support is available (+1 517 432 4998). The RDP-II staff may also be contacted via fax (+1 517 353 8957 Attn:RDP) or regular mail.

ACKNOWLEDGEMENTS

We thank several individuals for their past contributions: Robin Gutell (and his colleagues), Chuck Parker, Paul Saxman, Bonnie Maidak, Tim Lilburn, Niels Larsen, Tom Macke, Michael J. McCaughey, Ross Overbeek, Sakti Pramanik, Scott Dawson, Mitch L. Sogin, Gary Olsen and Carl Woese. The RDP is supported through US DOE OBER/NIH grant DE-FG02-99ER62848 and NSF grant DBI-0328255.

REFERENCES

1. Cole,J.R., Chai,B., Marsh,T.L., Farris,R.J., Wang,Q., Kulam,S.A., Chandra,S., McGarrell,D.M., Schmidt,T.M., Garrity,G.M. *et al*. (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*., **31**, 442–443.
2. Brown,M.P.S. (2000) Small subunit ribosomal RNA modeling using stochastic context-free grammar. In: *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000)*, San Diego, CA, pp. 57–66.

3. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *et al*. (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BioMed. Central Bioinformatics*, **3**, 2.
4. Garrity,G.M., Bell,J.A. and Lilburn,T.G. (2004) *Taxonomic Outline of the Prokaryotes. Bergey's Manual of Systematic Bacteriology*, 2nd edn. Release 5.0, May 2004. Springer-Verlag, NY.
5. Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L., Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **28**, 10–14.
6. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.